

Number of adaptive steps to a local fitness peak

KAVITA JAIN

*Theoretical Sciences Unit and Evolutionary and Organismal Biology Unit,
Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur P.O., Bangalore 560064, India*

PACS 87.23.-n – Ecology and evolution
PACS 02.50.-r – Probability theory, stochastic processes, and statistics
PACS 05.40.Fb – Random walks and Levy flights

Abstract. – We consider a population of genotype sequences evolving on a rugged fitness landscape with many local fitness peaks. The population walks uphill until it encounters a local fitness maximum. We find that the statistical properties of the walk length depend on whether the underlying fitness distribution has a finite mean. If the mean is finite, all the walk length cumulants grow with the sequence length but approach a constant otherwise. Experimental implications of our analytical results are also discussed.

Introduction. – The evolutionary process of adaptation is common in nature [1] and during the last decades, the dynamics of adaptation have been studied in several experiments on microbial populations [2]. The nature of the adaptive process depends crucially on the availability of beneficial mutations that improve the fitness [3]. If such mutations are readily available as in populations of very large size, the dynamics are well described by a deterministic theory [4] while for moderately large populations, a stochastic theory which accounts for competing multiple mutations can be applied [5]. Here we work in the parameter regime where beneficial mutations are rare and a population of genotype sequences performs an *adaptive walk* on a fitness landscape [6, 7].

More precisely, the adaptive walk model assumes that the number of mutants produced per generation is small so that the population is genetically homogeneous and may be represented by a single particle. The weak mutation assumption also renders the sequences differing by more than one mutation inaccessible. Furthermore the sequences carrying mutations that decrease the fitness do not survive and hence the adaptive walker always walks uphill. On a rugged fitness landscape with many local optima, the walk ends when a local fitness maximum is encountered since a better fitness is at least two mutations away as illustrated in Fig. 1. Remarkably, under these assumptions, the model depends only on a small set of parameters namely the sequence length and the fitness distribution underlying the fitness landscape. Recently some theoretical predictions for the first step [8] in the walk were tested in an experiment on a ssDNA virus [9] and a reasonable

agreement between theory and experiment was found. As the adaptive walk describes a simple and biologically realistic model of adaptation, it is important to analyse it in detail to extend our present understanding of adaptation dynamics.

In this Letter, we focus on the statistical properties of the length of adaptive walk defined as the number of beneficial mutations accumulated until the population reaches a local fitness maximum. Recently the walk length distribution was calculated within an approximation for the model described above [10] and the mean walk length was computed exactly in a simplified version of the adaptive walk [11]. However these studies assume that the fitness distribution has a finite mean. Here we relax this assumption and interestingly, we find that in the limit of infinitely long sequence, there is a transition in the behavior of the walk length distribution: it vanishes for fitness distributions with finite mean but remains finite otherwise. For finite sequences, this result implies that the walk length diverges with the sequence length for distributions with finite mean. For such distributions, we show that all the walk length cumulants grow logarithmically with the sequence length and find the proportionality constant for the first few cumulants. Our analytical results are compared with the numerical results and their experimental implications are also discussed.

Model. – We work with binary sequences of length L so that each sequence has L neighbors which are one mutation away. As the fitness always increases in an adaptive walk (see Fig. 1), the mutants that lower the current fitness h of the walker are rejected and a mutant with

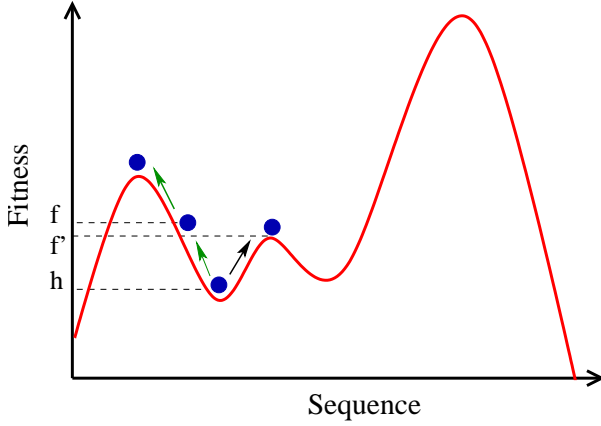


Fig. 1: (Color online) Schematic diagram to illustrate adaptive walk on a rugged fitness landscape with many local maxima. The population (filled circle) with fitness h has fitter one-mutant neighbors with fitness f, f', \dots . One of the better mutants is chosen with a transition probability (1). The global maximum is not accessible to the population as it is not a one-mutant neighbor and the walk terminates when the population reaches a local fitness maximum.

given fitness $f > h$ is chosen with a transition probability $T(f, h|f)$ proportional to the fitness difference $f - h$ [7]. Thus the normalised transition probability is given by

$$T(f, h|f) = \frac{f - h}{\sum_{g>h} g - h} \quad (1)$$

where the fitnesses are independent random variables chosen from a common distribution $p(f)$ with support on the interval $[0, u]$. Following previous works [11,12], we choose the fitnesses from a generalised Pareto distribution defined as

$$p(f) = (1 + \kappa f)^{-\frac{\kappa+1}{\kappa}} \quad (2)$$

where the fitness f is unbounded for $\kappa \geq 0$ and $f \leq -1/\kappa$ for $\kappa < 0$. The distribution of the *beneficial* mutations is however governed by the upper tail of the fitness distribution $p(f)$ [7] and hence can be one of the three universal distributions only [13,14]. The fitness distribution $p(f)$ lies in the domain of the extreme value distribution given by Weibull distribution for $\kappa < 0$, Gumbel distribution if $\kappa \rightarrow 0$ and Fréchet distribution for $\kappa > 0$. Although much of the experimental data on distribution of beneficial mutations is consistent with $\kappa \rightarrow 0$ [9,15], recent works also support $\kappa < 0$ [16] and $\kappa > 0$ [11].

The adaptive walk in the limits $\kappa \rightarrow \pm\infty$ is well studied theoretically. When $\kappa \rightarrow \infty$, the adaptive walk model reduces to a greedy walk [12] for which the walk length distribution is finite for infinitely long sequences [17] while for $\kappa \rightarrow -\infty$, a random adaptive walk is obtained [12] for which the walk length distribution is a Poisson distribution with mean $\ln L$ [18]. Recently the adaptive walk model described above was studied in detail for $\kappa = -1$ and $\kappa \rightarrow 0$ and the walk length distribution was computed

[10]. Here we are interested in the properties of adaptive walk when κ is arbitrary but finite.

Following [18], we consider the conditional probability $\mathcal{P}_J(f)$ that the walker takes at least J steps and has a fitness f at the J th step given that the initial fitness is f_0 . For long sequences, one can write down the following recursion relation for $J \geq 0$ [10]:

$$\mathcal{P}_{J+1}(f, L) = \int_0^f dh Lp(f)T(f, h|f) (1 - q^L(h))\mathcal{P}_J(h, L) \quad (3)$$

where $q(h) = \int_0^h dg p(g)$. The above equation expresses the fact that the walker can proceed to the next step if at least one fitness value greater than the current fitness h is available which occurs with a probability $1 - q^L(h) \approx 1 - e^{-L(1+\kappa h)^{-\frac{1}{\kappa}}}$. The walk length distribution Q_J that *exactly* J steps are taken is related to $\mathcal{P}_J(f)$ according to the following relation [10]:

$$Q_J(L) = \int_0^u dh q^L(h)\mathcal{P}_J(h, L) \quad (4)$$

This is because in order to terminate the walk at the J th step, none of the L mutant fitnesses at the next step should exceed the fitness at step J . In the following, we set the initial fitness f_0 to be zero, $\mathcal{P}_0(f, L) = \delta(f)$ which ensures that the walker does not start at a local fitness maximum.

Transition in the behavior of walk length. – Using a scaling analysis and extreme value theory, we now show that the qualitative behavior of walk length distribution Q_J changes at $\kappa = 1$. We find that the walk length distribution vanishes for $\kappa < 1$ but remains finite for $\kappa > 1$ as $L \rightarrow \infty$. We note that the behavior of Q_J discussed above for $\kappa \rightarrow \pm\infty$ is in accordance with our result.

For $\kappa < 1$, it is a good approximation to replace the sum on the right hand side (RHS) of (1) by the integral $L \int_h^u dg (g - h)p(g)$ when L is large [10]. Then in the limit $L \rightarrow \infty$, the recursion equation (3) reduces to

$$\bar{\mathcal{P}}_{J+1}(f) = (1 - \kappa) \int_0^f dh \frac{(f - h)p(f)}{(1 + \kappa h)^{\frac{\kappa-1}{\kappa}}} \bar{\mathcal{P}}_J(h) \quad (5)$$

where $\bar{\mathcal{P}}_J(f) \equiv \mathcal{P}_J(f, L \rightarrow \infty)$. A generating function for the distribution $\bar{\mathcal{P}}_J(f)$ can be calculated (see (15)) which shows that $\bar{\mathcal{P}}_J(f)$ is finite. Thus from (4), it immediately follows that $Q_J(L) \rightarrow 0$ as $L \rightarrow \infty$ for all J . Our numerical results in Fig. 2 for $\kappa = 1/2$ show that for $J > 4$, the distribution $Q_J(L = 10^4) < Q_J(L = 10^3)$ and for $J < 4$, $Q_J(L = 10^5) < Q_J(L = 10^4)$. Thus the distribution Q_J decreases with increasing L .

For $\kappa \geq 1$, the sum in (1) can not be replaced by an integral as the mean of the distribution is infinite. For such fat-tailed distributions, the sum of L random variables is dominated by the largest value \tilde{f} amongst them [13,14]. If at most one fitness exceeds \tilde{f} , we have $L(1 - q(\tilde{f})) \sim 1$ or $1 + \kappa\tilde{f} = L^\kappa$ for any κ . Using this result in the recursion

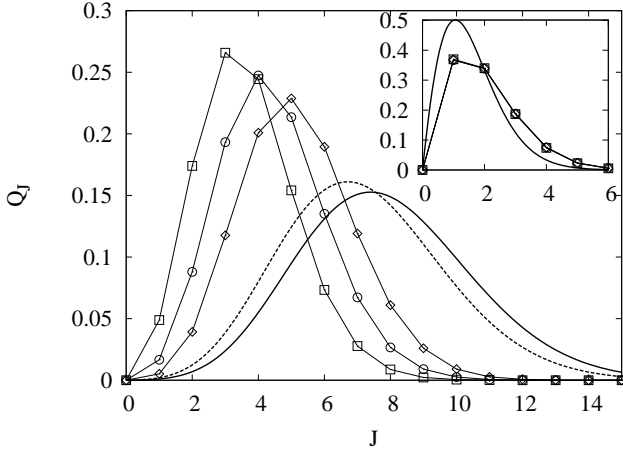


Fig. 2: Walk length distribution Q_J as a function of J for $\kappa = 1/2$ (main) and $3/2$ (inset) for $L = 10^3$ (squares), 10^4 (circles) and 10^5 (diamonds) to show that in the infinite sequence length limit, Q_J vanishes for $\kappa < 1$ but remains finite for $\kappa > 1$. The points are joined to guide the eye. For comparison, the analytical result for the random adaptive walk (main) is shown for $L = 500$ (broken line) and 10^3 (solid line) and for the greedy walk (inset) for infinitely long sequence.

equation (3) and changing the variable to $z = (1 + \kappa f)/L^\kappa$, we find that for $\kappa > 1$,

$$\mathcal{P}_{J+1}(z, L) \propto \int_{L^{-\kappa}}^z dy (z - y) z^{-\frac{\kappa+1}{\kappa}} (1 - e^{-y^{-\frac{1}{\kappa}}}) \mathcal{P}_J(y, L) \quad (6)$$

where the proportionality constant depends on κ and is omitted for brevity. Since the distribution $\mathcal{P}_1(f, L)$ for large L is writeable as

$$\mathcal{P}_1(f, L) \approx \frac{\kappa - 1}{L^\kappa} \left(\frac{1 + \kappa f}{L^\kappa} \right)^{-1/\kappa} = \frac{\kappa - 1}{L^\kappa} z^{-1/\kappa} \quad (7)$$

it follows that for large L , the fitness distribution at the J th step of adaptive walk is of the following scaling form:

$$\mathcal{P}_J(f, L) \approx \frac{1}{L^\kappa} S_J \left(\frac{1 + \kappa f}{L^\kappa} \right) \quad (8)$$

where $S_J(z)$ is a scaling function. Using this scaling form in (4) and taking the limit $L \rightarrow \infty$, we immediately find that $Q_J \approx (1/\kappa) \int_0^\infty dz S_J(z) e^{-z^{-\frac{1}{\kappa}}}$ is finite in agreement with the numerical results shown in the inset of Fig. 2.

Walk length cumulants for fitness distributions with finite mean. — For $\kappa < 1$, the probability that the walk terminates at the J th step is zero or in other words, the walk goes on indefinitely for infinitely long sequences and hence the mean number of adaptive steps diverges with L . We now show that all the walk length cumulants increase logarithmically with L .

On differentiating (3) twice with respect to f and writing $\mathcal{P}_J(f, L) = p(f)P_J(f, L)$, a straightforward calculation

shows that the distribution $P_J(f, L)$ obeys the following equation [10]:

$$P''_{J+1}(f, L) = \frac{p(f)(1 - q^L(f))}{\int_f^u dg (g - f)p(g)} P_J(f, L), \quad J \geq 1 \quad (9)$$

where prime denotes a f -derivative. The boundary conditions are given by [10]

$$P_J(0, L) = 0, \quad P'_J(0, L) = \frac{\delta_{J,1}}{\int_0^u dg g p(g)} \quad (10)$$

As (9) is non-diagonal in J , we work with a generating function $G(x, f) = \sum_{J=1}^\infty P_J(f)x^J$, $x < 1$ which obeys the following second order differential equation:

$$G''(x, f) = \frac{x p(f)(1 - q^L(f))}{\int_f^u dg (g - f)p(g)} G(x, f) \quad (11)$$

The above differential equation does not appear to be exactly solvable due to the factor $1 - q^L(f)$ on the RHS. As this cumulative probability decreases from one to zero with increasing f , we consider (11) by approximating

$$1 - q^L(f) = 1 - e^{-\left(\frac{1+\kappa f}{1+\kappa \tilde{f}}\right)^{-\frac{1}{\kappa}}} \approx \begin{cases} 1, & f < \tilde{f} \\ r(f), & f > \tilde{f} \end{cases} \quad (12a)$$

where $1 + \kappa \tilde{f} = L^\kappa$ as found earlier. Equation (11) has been solved by choosing $r(f) = 0$ in [10] for $\kappa = -1$ and $\kappa \rightarrow 0$. Here we show that the leading order behavior of the cumulants does not depend on the choice of $r(f)$. For $f < \tilde{f}$, as a result of (12a), we have

$$G''_{<} = \frac{x(1 - \kappa)}{(1 + \kappa f)^2} G_{<} \quad (13)$$

whose solution is of the form $G_{<} = a_+(1 + \kappa f)^{\alpha_+} + a_-(1 + \kappa f)^{\alpha_-}$ where

$$\alpha_{\pm}(x) = \frac{1 \pm \sqrt{1 + \frac{4x(1 - \kappa)}{\kappa^2}}}{2} \quad (14)$$

and the constants a_{\pm} can be determined using the boundary conditions (10) to finally yield

$$G_{<} = \frac{x(1 - \kappa)}{\sqrt{\kappa^2 + 4x(1 - \kappa)}} [(1 + \kappa f)^{\alpha_+} - (1 + \kappa f)^{\alpha_-}] \quad (15)$$

For $f > \tilde{f}$, using (12b) in (11), we get

$$G''_{>} = \frac{x(1 - \kappa)r(f)}{(1 + \kappa f)^2} G_{>} \quad (16)$$

whose solution is of the form

$$G_{>} = b_1 g_1(x, f) + b_2 g_2(x, f) \quad (17)$$

where the functions g_1, g_2 obey (16) and b_1, b_2 are constants. In order to compute the walk length cumulants for

large L , it is sufficient to find the L dependence of b_1, b_2 . This can be done by matching the solutions $G_<$ and $G_>$ and their first derivative at $f = \tilde{f}$ and we find

$$b_1(x, L) = L^{\kappa\alpha+} b_{11}(x) + L^{\kappa\alpha-} b_{12}(x) \quad (18)$$

$$b_2(x, L) = L^{\kappa\alpha+} b_{21}(x) + L^{\kappa\alpha-} b_{22}(x) \quad (19)$$

where $b_{ij}(x)$ are independent of L .

We now use (15) and (17) to write down an expression for the generating function $H(x, L) = \sum_{J=1}^{\infty} Q_J(L) x^J$ of the walk length distribution. Using (12a) in (4), we have

$$H(x, L) \approx \int_{\tilde{f}}^u dh (1 - r(h)) p(h) G_>(x, h) \quad (20)$$

$$\propto \int_1^{\frac{1+\kappa u}{L^\kappa}} \frac{dz}{L} z^{-\frac{\kappa+1}{\kappa}} (1 - r(z)) G_>(x, z) \quad (21)$$

$$= L^{\kappa\alpha+ - 1} T_1(x) + L^{\kappa\alpha- - 1} T_2(x) \quad (22)$$

where the integral

$$T_i(x) \propto \int_1^{\frac{1+\kappa u}{L^\kappa}} dz z^{-\frac{\kappa+1}{\kappa}} (1 - r(z)) (b_{1i} g_1(z) + b_{2i} g_2(z))$$

is independent of L which can be seen using the upper bounds namely $u = -1/\kappa$ for $\kappa < 0$ and infinity for $\kappa \geq 0$. Since $\kappa\alpha_{\pm} - 1 < 0$ for any κ , on taking the limit $L \rightarrow \infty$ in (22), we find that the walk length distribution vanishes as discussed earlier.

The fact that $T_i(x)$ is independent of L leads to a considerable simplification of the problem and allows us to find the cumulants to leading order in sequence length. The n th cumulant is defined as [14]

$$c_n(L) = \left. \frac{d^n \ln H(x, L)}{ds^n} \right|_{s=0} \quad (23)$$

where $s = \ln x$. As the first term on the RHS of (22) decays less rapidly than the second term for any κ , we have $H(x) \approx L^{\kappa\alpha+ - 1} T_1(x)$. Using this, we immediately obtain the cumulants to leading order in L as

$$c_n \approx \frac{\ell}{2} \frac{d^n}{ds^n} \sqrt{\kappa^2 + 4e^s(1 - \kappa)} \Big|_{s=0}, \quad n > 0 \quad (24)$$

where $\ell = \ln L$. Thus we find that all the walk length cumulants increase logarithmically with L . The first three cumulants computed using the last expression are given by

$$c_1 \approx \frac{1 - \kappa}{2 - \kappa} \ell \quad (25)$$

$$c_2 \approx \frac{(1 - \kappa)(2 - 2\kappa + \kappa^2)}{(2 - \kappa)^3} \ell \quad (26)$$

$$c_3 \approx \frac{(1 - \kappa)(4 - 8\kappa + 6\kappa^2 - 2\kappa^3 + \kappa^4)}{(2 - \kappa)^5} \ell \quad (27)$$

In the limit $\kappa \rightarrow -\infty$, all the above cumulants are equal to $\ln L$ in agreement with the results for random adaptive

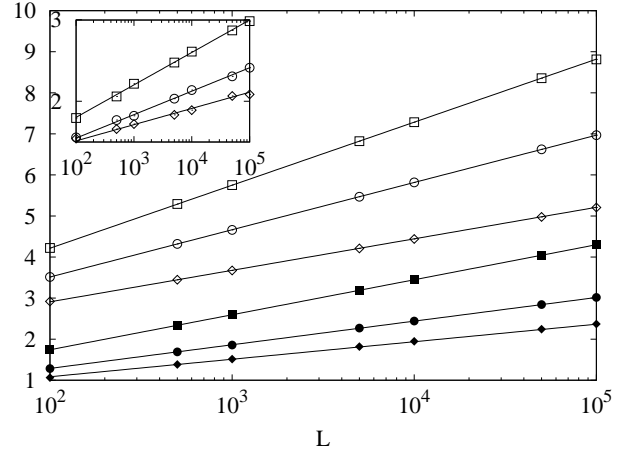


Fig. 3: Plot of the first three cumulants as a function of sequence length L for fitness distributions $p(f) = 1$ (squares), e^{-f} (circles) and $(1 + 0.5f)^{-3}$ (diamonds). The main figure shows the simulation data for mean c_1 (open symbols) and variance c_2 (filled symbols) and the inset shows the third cumulant c_3 . The slope of the solid lines is given by the analytical results in (25)-(27). The numerical data has been averaged over 10^6 independent realisations of fitnesses and the data for c_2 has been shifted by a constant for clarity.

walk [18]. We also recover the previous results for uniformly and exponentially distributed fitnesses [10]. Equations (25) and (26) also match the results of [11] in which a fixed set of mutants during the entire walk is assumed. In contrast, we have considered a more realistic mutation scheme in which a novel set of mutants are available to the population at each adaptive step. The above expressions for c_1 and c_2 have also been seen in a deterministic model of evolution [19] and a relationship of this model to adaptive walks has been recently elucidated [11]. Figure 3 shows that our expressions (25)-(27) agree very well with the numerical results.

Discussion. — In this article, we studied a biologically realistic model of adaptation and showed that to leading orders in L , the average walk length is a constant for fitness distributions with infinite mean but increases logarithmically with the sequence length otherwise. Our analytical results agree well with the numerical simulations.

Our broad theoretical result that the adaptive walks are short (see Fig. 3) is consistent with the experiments on microbes [2] and fungus [20] in which 2 – 6 adaptive substitutions have been observed. However more detailed experimental studies are needed to test our predictions. Our result (25) shows that the walk should last longer in systems with smaller κ . This may be checked by measuring the mean walk length in populations with $\kappa = -1$ [16], $\kappa \rightarrow 0$ [15] and $\kappa > 0$ [11]. To find the dependence of walk length properties on L , varying the sequence length may not be experimentally viable but it should be possi-

ble to set up experiments along the lines of [9] and vary the initial fitness rank. If the initial ranks are of the order L , we expect our analysis to hold [8]. Experimental data for the walk length distribution showing insensitivity to the initial rank would then imply an underlying fat-tailed fitness distribution with infinite mean.

Acknowledgement. – The author thanks J. Krug for useful comments on the manuscript.

REFERENCES

- [1] H. A. Orr. *Nat. Rev. Genet.*, 6:119–127, 2005.
- [2] S.F. Elena and R.E. Lenski. *Nat. Rev. Genet.*, 4:457–469, 2003.
- [3] K. Jain and J. Krug. *Genetics*, 175:1275, 2007; K. Jain, J. Krug, and S.-C. Park. *Evolution*, 65:1945, 2011.
- [4] K. Jain and J. Krug. *J. Stat. Mech.: Theor. Exp.*, page P04008, 2005; D.B. Saakian and C.-K. Hu. *Proc. Natl. Acad. Sci. USA*, 103:4935–4939, 2006.
- [5] P.J. Gerrish and R.E. Lenski. *Genetica*, 102:127–144, 1998; S.-C. Park and J. Krug. *Proc. Natl. Acad. Sci. USA*, 98:18135–18140, 2007.
- [6] J. Maynard Smith. *Nature*, 225:563, 1970.
- [7] J. H. Gillespie. Oxford University Press, Oxford, 1991.
- [8] H. A. Orr. *Evolution*, 56:1317–1330, 2002; H. A. Orr. *Evolution*, 60:1113, 2006.
- [9] D.R. Rokyta, P. Joyce, S.B. Caudle, and H.A. Wichman. *Nat. Genet.*, 37:441–444, 2005.
- [10] K. Jain and S. Seetharaman. [arXiv:1104.5583](#) (to appear in *Genetics*)
- [11] J. Neidhart and J. Krug. [arXiv:1105.0592](#) (to appear in *Phys. Rev. Lett.*)
- [12] P. Joyce, D.R. Rokyta, C. J. Beisel and H.A. Orr. *Genetics*, 180:1627–1643, 2008.
- [13] J.-P. Bouchaud and A. Georges. *Phys. Rep.*, 195:127–293, 1990.
- [14] D. Sornette. Springer, Berlin, 2000.
- [15] A. Eyre-Walker and P.D. Keightley. *Nat. Rev. Genet.*, 8:610, 2007.
- [16] D.R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H.A. Wichman. *J Mol Evol*, 69:229, 2008.
- [17] H. A. Orr. *J. theor. Biol.*, 220:241–247, 2003.
- [18] H. Flyvbjerg and B. Lautrup. *Phys. Rev. A*, 46:6714–6723, 1992.
- [19] C. Sire, S.N. Majumdar and D.S. Dean. *J. Stat. Mech.*, L07001, 2006.
- [20] S.E. Schoustra, T. Bataillon, D.R. Gifford and R. Kassen. *PLoS Biol.*, 7:e1000250, 2009.